The Effect of Alignment Correction on Cross-Lingual Annotation Projection

Shabnam Behzad*[1], Seth Ebner*[2], Marc Marone[2], Benjamin Van Durme[2], Mahsa Yarmohammadi[2]

[1]Georgetown University,[2]Johns Hopkins University

{seth,mmarone1,vandurme,mahsa}@jhu.edu,
shabnam@cs.georgetown.edu

Abstract

Cross-lingual annotation projection is a practical method for improving performance on low resource structured prediction tasks. An important step in annotation projection is obtaining alignments between the source and target texts, which enables the mapping of annotations across the texts. By manually correcting automatically generated alignments, we examine the impact of alignment quality—automatic, manual, and mixed—on downstream performance for two information extraction tasks and quantify the trade-off between annotation effort and model performance.

1 Introduction

Cross-lingual annotation projection (Yarowsky and Ngai, 2001) involves mapping source annotations to target text through alignments. Recent studies such as Yarmohammadi et al. (2021) and Chen et al. (2022) suggest that word alignment quality substantially impacts downstream performance.

Automatic word alignments are inexpensive to obtain but may be of low quality. On the other side of the alignment quality spectrum (Figure 1) are manual (human-labeled) alignments, which are expensive but accurate. Our goals are to quantify the impact of automatic vs. manual alignments on downstream task performance and to explore the quality spectrum to quantify the trade-off between automatic and manual alignments in terms of downstream performance and cost.

We investigate the alignment quality spectrum on two structured prediction tasks: shallow semantic parsing (BETTER Basic IE[1]) and named entity recognition (NER). In the BETTER IE scenario, we start with a typical fully automatic *translate, align, and project* pipeline, so-called "silver" data creation, and compare with manually labeled

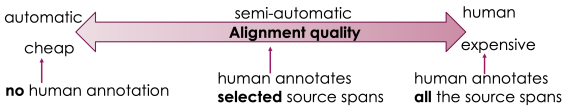


Figure 1: Alignment annotation spectrum. It is often easier to annotate alignments than to train annotators for a complex downstream task. Higher quality alignments improve performance but are more costly.

data (annotated by two of the authors). For NER, the dataset we use already has annotations in two languages (with gold bitext and gold word alignments).

Rather than manually annotate alignments for every example, which is expensive as shown on the right end of the spectrum in Figure 1, we collect manual annotations on the data for which two alignment methods—silver data creation and an unsupervised embedding-based span alignment tool—point to different target spans (§5). Our contributions are: 1) evidence that manually correcting alignments improves downstream performance, 2) evidence that downstream performance correlates with the amount of manual alignment effort, and 3) analysis on the types of spans that are manually corrected.

2 Related Work

Cross-lingual projection is a method of transferring annotations from a source language to a target language that has often been used to increase performance on the target language (Yarowsky and Ngai, 2001; Yarmohammadi et al., 2021; Chen et al., 2022, *i.a.*), but its utility depends on obtaining reliable alignments between the source and target text. There has been extensive research on supervised and unsupervised alignment at the word, phrase, and sentence levels (Zhang et al., 2016; Jalili Sabet et al., 2020; Nagata et al., 2020; Chousa et al., 2020; Chen et al., 2021; Li et al., 2022, *i.a.*). However, these studies report primarily intrinsic evaluation of alignment quality and leave extrinsic

*Equal contribution
[1]https://ir.nist.gov/better

Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 244–251

	BETTER (%)	NER (%)
Unique source spans	4896	3180
Identical automatic spans	2432 (50%)	1942 (61%)
Candidates for correction	2464 (50%)	1238 (39%)

Table 1: Number of source spans in total and in the candidate subset for correction. Candidates are spans that are non-identical (different or overlapping), based on two automatic alignment results. BETTER candidate spans are shown to humans for re-alignment. NER candidate spans are corrected according to gold alignments from the original resource.

evaluation on downstream tasks underexplored.

Stengel-Eskin et al. (2019) showed that gold alignment data has a greater impact on word alignment performance than the amount of pre-training bitext does, suggesting that the benefits of manually correcting alignments may also extend to the creation of higher-quality projected data.

Our work responds to these lines of prior work by examining the extrinsic downstream impact of our approaches and how the incorporation of different amounts of gold alignment data affects performance. In contrast to prior work on projection, such as Yarmohammadi et al. (2021), our work focuses on the impact of the alignment component of the projection pipeline on the overall effectiveness of cross-lingual projection and explicitly adjusts automatic alignments with manual corrections.

3 Methods

Fully Automatic We consider improving the zero-shot learning scenario where the gold training data is in a different language than the target data we want to evaluate on. For both tasks, we explore multiple setups that include training on gold English data alone (zero-shot) or combined with projected target-language data. Projected data is created by transferring the gold source labels to translated target text via automatically obtained or manually corrected word alignments.

Silver Data Creation To create silver data, we followed the process of Yarmohammadi et al. (2021). First, if there was no gold translation of the source text (as in BETTER), we translated the source text into the target language using a state of the art translation system (Xu et al., 2021). Second, we obtained word alignments between the original and target parallel text using `awesome-align` (Dou and Neubig, 2021), a state of the art contextualized embedding-based word aligner (see Appendix A for further details). Finally, we projected the annotations from the source language to the target language based on the word alignments. For multi-word spans, the target span is a contiguous span containing all aligned words from the same source span.

Unsupervised Span Alignment We implemented a span alignment tool by extending the techniques of SimAlign (Jalili Sabet et al., 2020) to compute similarities between the representations of spans rather than of tokens.[2] We used a frozen pre-trained encoder and did not update any model parameters. We performed a hyperparameter search (Appendix C) to configure an aligner that most frequently produces spans identical to those of `awesome-align`.

Alignment Correction After obtaining "silver spans" from `awesome-align` and "unsupervised spans" from the span aligner, we selected source spans which the two methods aligned to different target spans. Around half the total source spans were selected for re-alignment (Table 1). For BETTER, we asked human annotators to re-align selected spans from scratch. For NER, we simulated manually correcting automatic alignments by retrieving gold alignments from the original manually annotated resource.[3] Further details are given in §5. We refer to data projected after the alignment correction step as *semi-automatic* because it is created from a mix of automatic and manual alignments.

4 Tasks

We investigate the impact of alignment quality using established models, as modeling improvements are outside the scope of this work.

4.1 NER

Data We utilized GALE (Li et al., 2015), which includes word-aligned Chinese and English parallel text. Alignments were obtained from multiple rounds of human annotations. As a part of the OntoNotes corpus (Weischedel et al., 2013), a portion of the Chinese section of GALE was annotated for NER. The gold-aligned NER data consists of 2385, 287, and 189 sentences in the train, dev, and

[2]We considered all spans of up to a certain length, giving a linear number of spans per sentence.

[3]To avoid issues with labeling overlapping spans in the BIO-tagged NER data, we use the gold alignments for the entire sentence rather than for individual spans. In many cases, identical `awesome-align` silver spans and unsupervised spans present in the sentence that should not have been re-aligned closely match the gold-aligned spans anyway.

test splits, respectively. Further human annotation was not necessary for this task because data and annotations were already available in both languages.

Model We used a BERT-based token tagging NER model,[4] which outputs tag probabilities via a softmax on logits from a linear layer on top of a bert-base-multilingual-cased encoder (Devlin et al., 2019). We select the checkpoint from the epoch that achieved the best F1 performance on the gold Chinese dev set. We evaluate our models using micro-averaged F1.

4.2 BETTER

Data The Better Extraction from Text Towards Enhanced Retrieval (BETTER) Program[5] develops methods for event extraction in a target language, given gold annotations only in English. We focus on the "Basic" events level, where the goal is to identify events and their arguments (agents and patients), i.e. shallow semantic parsing, with Farsi as the target language.

Model Our BETTER IE system is based on the Spanfinder model (Xia et al., 2021), consisting of a contextualized encoder and a BiLSTM-CRF span tagger. The model first extracts event anchors and labels them with event types. Conditioned on an anchor span, the model then identifies argument spans (agents and patients). We report the program-defined "combined F1" metric, which is the product of "event match F1" and "argument match F1" based on an alignment of predicted and reference event structures.

Annotation Task We gather alignment corrections through the TASA[6] human annotation interface (Stengel-Eskin et al., 2019). Additional information about the interface is given in Appendix B. 2,464 candidate source spans (training and analysis) occur across 1,012 sentence pairs. Each sentence pair, containing one or more source spans highlighted for alignment, is considered a task. Tasks took 1 minute on average for a total annotation time of ~ 16 hours.

5 Experiments and Results

We compare the performance of models trained on various combinations of gold, silver, and semi-automatic training data. We evaluate the models on the target language test sets: the Chinese gold test set for NER and Farsi semi-automatic analysis set for BETTER.

We also use English training data in two different ways: combined with the target language data ('En +' in Table 2 and Table 3) and as pre-training before we fine-tune on the target data ('En →' in Table 2).

Training Data	Micro-F1 on Gold Zh Test Set
En (zero-shot baseline)	17.6
Gold Zh (upper bound)	74.7
Silver Zh	44.0
Semi-automatic Zh	56.0
Gold projection Zh	58.9
En + Silver Zh	35.1
En + Semi-automatic Zh	51.9
En + Gold projection Zh	60.7
En → Silver Zh	45.5
En → Semi-automatic Zh	53.8
En → Gold projection Zh	51.7

Table 2: NER results on GALE Chinese gold test set. In general, as alignment quality increases, downstream performance increases.

5.1 NER

The results in Table 2 show that by augmenting the source language training data (En) with data in the target language (Zh), zero-shot performance can be much improved. When projection is performed via gold word alignments (Gold projection Zh), there is 15.8% absolute performance degradation compared to when gold Chinese data (Gold Zh) is used for training. Thus, there is some loss in performance when using projected data. Training on alignment-corrected (semi-automatic) data outperforms training on silver Chinese data in all settings. Performance per entity type in a representative experimental setting is shown in Table 5. Overall, performance correlates with the amount of manual effort used in creating the data.

5.2 BETTER

The results in Table 3 show that using projected training data, either by itself or combined with source language (En) training data, outperforms the zero-shot setting. Even though the majority of the semi-automatic spans match the silver spans (see §6.2 for details), replacing the silver data with the semi-automatic data improves the BETTER scores. However, the gain is not as substantial as that for NER. This could be due to the BETTER an-

[4] Adapted from http://github.com/kamalkraj/BERT-NER and Stengel-Eskin et al. (2019).

[5] https://www.iarpa.gov/index.php/research-programs/better

[6] https://github.com/hltcoe/tasa

Train	Test on Semi-automatic Fa
En (zero-shot baseline)	36.9
Silver Fa	40.1
Semi-automatic Fa	40.7
En + Silver Fa	44.8
En + Semi-automatic Fa	45.1

Table 3: BETTER results on semi-automatic Farsi analysis set.

alysis data being non-gold or the BETTER scorer's sensitivity to small changes in predictions.

6 Analysis

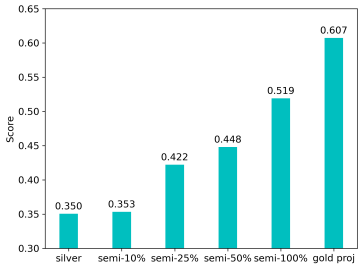
6.1 Annotation Budget Constraints

Annotation budgets constrain how much of the data can be projected along gold alignments. Furthermore, it is cheaper to annotate alignments than to train annotators on a complex task. Note that our annotation task took only ~ 16 hours for all data. We analyze downstream performance as a function of the amount of data correction.

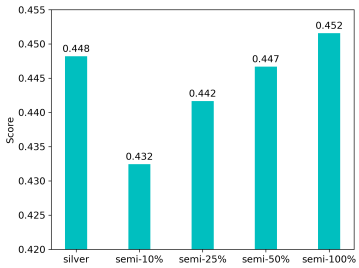
For both NER and BETTER, we subsample 10%, 25%, and 50% of the tasks (sentences) in which the two automatic alignment methods have a disagreement, and then replace their silver alignments with the correct alignments: gold (for NER) and manually corrected (for BETTER). The remaining portion of the data comes from the silver data aligned with `awesome-align` (i.e., semi-10% consists of 10% corrected data and 90% automatic silver data). We do the sampling process 5 times and report the average performance.

The results in Figure 2a show that the performance consistently improves as subsampling percentage increases. In all of the cases where projected Zh is combined with En, either added or fine-tuned, any percentage of manual correction outperforms training on silver Zh only (see Appendix E for the plot for the fine-tuning setting).

Figure 2b shows a similar trend for BETTER. The plots suggest, however, that to outperform training on silver data, more than 50% of the alignments need to be corrected. We hypothesize this is in part due to the evaluation set being non-gold, so correcting the training data shifts away from the distribution of the evaluation data.



(a) NER: Micro-F1 on gold Chinese test set



(b) BETTER: Combined F1 on semi-automatic (corrected) Farsi analysis set

Figure 2: Performances for (a) NER and (b) BETTER for systems trained on gold English combined with: automatic (silver), subsampled semi-automatic (semi-x%), or gold projection (gold proj) data.

6.2 Extent of Alignment Disagreements

Around 86% and 60% of all manually corrected BETTER target spans match, either fully or partially, the automatic silver spans or the automatic unsupervised spans, respectively. We consider two spans to be partially matching or overlapping if the length of the longest consecutive common character sequence is larger than 30% of the length of the longer span. As Table 4 shows, most of the silver and manually annotated spans are identical (73.4%), whereas the unsupervised and manually annotated spans mostly are overlapping (52.4%).

6.3 Alignment Agreement Per Span Type

Table 4 shows that there is no substantial difference between the types of spans that have been categorized as identical, overlapping, or different. The silver and unsupervised span alignment approaches do not seem to do particularly better or worse on event anchors, agents, or patients.

For the NER task we observe that overall, the silver span extraction setup gave better quality spans compared to the unsupervised span alignment approach. See Appendix F for results per entity type.

	Identical Spans				Overlapping Spans				Different Spans			
	Anchor	Agent	Patient	Total (%)	Anchor	Agent	Patient	Total (%)	Anchor	Agent	Patient	Total (%)
Silver	669	679	473	73.4	163	75	73	12.5	117	107	126	14.1
Unsupervised	59	66	64	7.6	470	450	380	52.4	420	345	228	40.0

Table 4: (Dis)agreement of semi-automatic spans with automatic spans in BETTER.

7 Conclusion and Future Work

In this paper, we investigated how much the quality of alignments impacts downstream task performance when annotations are projected from English to another language. Our experimental results show that the utilization of different alignment methodologies, followed by corrections of the disagreements arising from such approaches, can reduce human effort while improving results.

There are several promising avenues for future research in cross-lingual annotation projection. Although this study did not investigate the influence of translation quality on downstream performance (which is the first step of the data projection pipeline), we believe this could yield important findings. Another direction involves the investigation of active learning techniques for prioritizing which alignments to correct.

Limitations

The automatic data creation procedure can introduce errors during both the translation and alignment steps. This study is limited to the errors during the alignment step. Even though we used state of the art machine translation techniques in this work, the translation errors could still affect the quality of the alignments or projected data, especially in silver data creation.

Moreover, we studied only two tasks, NER and BETTER, and considered only two target languages, Chinese and Farsi. Tasks with significantly different structures (e.g., deep parsing) may be affected differently by alignment corrections.

Acknowledgments

We acknowledge support through a fellowship from JHU + Amazon Initiative for Interactive AI (AI2AI). This work was supported in part by IARPA BETTER (#2019-19051600005). The views and conclusions contained in this work are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, or endorsements of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Chi Chen, Maosong Sun, and Yang Liu. 2021. Mask-align: Self-supervised neural word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4781–4791, Online. Association for Computational Linguistics.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2022. Frustratingly easy label projection for cross-lingual transfer. *arXiv preprint arXiv:2211.15613*.

Katsuki Chousa, Masaaki Nagata, and Masaaki Nishino. 2020. SpanAlign: Sentence alignment method based on cross-language span prediction and ILP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4750–4761, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics:*

EMNLP 2020, pages 1627–1643, Online. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lei Li, Kai Fan, Hongjia Li, and Chun Yuan. 2022. Structural supervision for word alignment and machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4084–4094, Dublin, Ireland. Association for Computational Linguistics.

Xuansong Li, Stephen Grimes, Stephanie Strassel, Xiaoyi Ma, Nianwen Xue, Mitch Marcus, and Ann Taylor. 2015. Gale chinese-english parallel aligned treebank–training. *Linguistic Data Consortium, Philadelphia, PA*.

Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. 2020. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, Online. Association for Computational Linguistics.

Elias Stengel-Eskin, Tzu-ray Su, Matt Post, and Benjamin Van Durme. 2019. A discriminative neural model for cross-lingual word alignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 910–920, Hong Kong, China. Association for Computational Linguistics.

Leila Tavakoli and Heshaam Faili. 2014. Phrase alignments in parallel corpus using bootstrapping approach. *International Journal of Information and Communication Technology*, 6:63–76.

Shubham Toshniwal, Haoyue Shi, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel. 2020. A cross-task analysis of text span representations. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 166–176, Online. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. LOME: Large ontology multilingual extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, Online. Association for Computational Linguistics.

Haoran Xu, Benjamin Van Durme, and Kenton Murray. 2021. BERT, mBERT, or BiBERT? a study on contextualized embeddings for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6663–6675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, Aaron Steven White, Mark Dredze, and Benjamin Van Durme. 2021. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Dongxu Zhang, Boliang Zhang, Xiaoman Pan, Xiaocheng Feng, Heng Ji, and Weiran Xu. 2016. Bitext name tagging for cross-lingual entity annotation projection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 461–470, Osaka, Japan. The COLING 2016 Organizing Committee.

A `awesome-align` Hyperparameters

We used `awesome-align` (Dou and Neubig, 2021), a contextualized embedding-based word aligner that extracts word alignments based on token embedding similarities. We fine-tuned the underlying XLM-R encoder on around two million parallel sentences from the OSCAR corpus (Abadji et al., 2022) of English-Farsi and English-Chinese pairs. We further fine-tuned the encoder for BETTER on English-Farsi gold alignments on 1500 sentence pairs by Tavakoli and Faili (2014). We reused empirically-chosen `awesome-align` hyperparameters from prior work for a similar task (Yarmohammadi et al., 2021): softmax normalization with probability thresholding of 0.001, 4 gradient accumulation steps, 1 training epoch with a learning rate of 2×10^{-5}, alignment layer of 16, and masked language modeling ("mlm"), translation language modeling ("tlm"), self-training objective ("so"), and parallel sentence identification ("psi") training objectives. We further fine-tuned the resulting model on the gold word alignments with the same hyperparameters, for 5 training epochs

with a learning rate of 10^{-4} and only "so" as the training objective.

B Annotation Interface

In each task, a pair of tokenized sentences, one in English (source) on the top and one in Farsi (target) on the bottom are shown to the user. In each English sentence, there are one or more spans to align, as highlighted in Figure 3. The user needs to annotate the English spans word by word.

C Unsupervised Span Alignment Hyperparameters

We did a random search to find the best hyperparameters for the unsupervised span aligner. We selected the following for NER:

- Span enumeration strategy: full (all spans of length up to maximum width)
- Max target span width: 5 tokens
- Alignment decoding method: greedy (decode alignments in decreasing order of similarity score)
- Allow overlap: true (alignments can contain overlapping spans)
- Span representation: diff-sum (Toshniwal et al., 2020)
- Encoder and layer: bert-base-multilingual-cased (Devlin et al., 2019), layer 7 (0-indexed)
- Coupled: false (encode source and target text separately)

and the following for BETTER:

- Span enumeration strategy: full (all spans of length up to maximum width)
- Max target span width: 4 tokens
- Alignment decoding method: greedy (decode alignments in decreasing order of similarity score)
- Allow overlap: true (alignments can contain overlapping spans)
- Span representation: endpoint (concatenate embeddings of first and last subtokens)
- Encoder and layer: EnFav1.0 (internal bilingual model), layer 15 (0-indexed)
- Coupled: true (encode source and target text as a single sequence)

D NER Model Hyperparameters

- Encoder: bert-base-multilingual-cased (Devlin et al., 2019)
- Max sequence length: 128 WordPieces
- Batch size: 32
- Optimizer: Adam (Kingma and Ba, 2014)
- Learning rate: 5×10^{-5}
- Learning rate linear warmup: 10% of training steps
- Epochs: 25

E Annotation Budget Constraints in Fine-tuning Setting

Figure 4 shows the results of semi-automatic data subsampling experiments for NER when the models are pre-trained on English and fine-tuned on Chinese data. The performance improves with increasing subsampling percentage. However, there is a slight degradation when using the entire semi-automatic data compared to 50%. Gold projection is unexpectedly lower than the semi-50% and semi-100% settings.

Experiments for BETTER in which we train on only Farsi data, not combined with gold English, show inconsistent results. We hypothesize this is due to training and validating on noisy non-gold data. Further investigations of these phenomena are left as future work.

F Alignment Agreement Per Entity Type

The entity types that needed the most correction in both extraction settings were NORP (nationalities or religious or political groups) and PERCENT (percentage, including "%"). PERSON (people, including fictional), MONEY (monetary values, including unit), and WORK-OF-ART (titles of books, songs, etc.) were among the entity types with the highest number of identical spans for both the silver span extraction and unsupervised approaches (Figure 5).

Figure 3: An example of the annotation interface for BETTER.

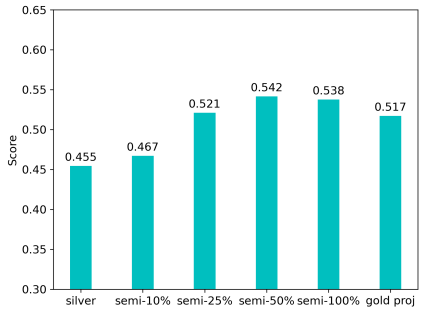
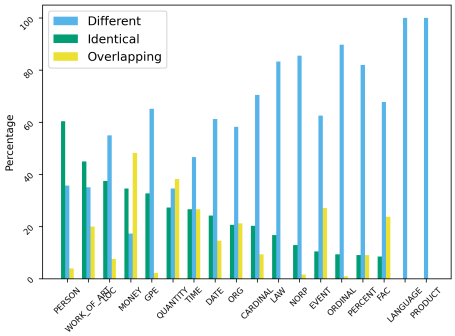


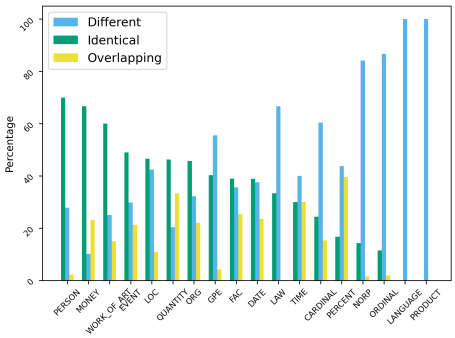
Figure 4: NER Micro-F1 on gold Chinese test set. Models are pre-trained on English data and fine-tuned on projected Chinese data.

Type	# Train	# Test	F1 (%)
GPE	921	41	75.6
CARDINAL	440	17	50.0
DATE	351	24	63.6
ORG	335	32	47.4
NORP	256	17	37.5
PERSON	230	31	51.5
MONEY	139	6	100.0
ORDINAL	107	6	20.0
PERCENT	100	0	NA
LOC	80	9	31.6
FAC	59	7	0.0
QUANTITY	55	3	57.1
EVENT	48	4	50.0
TIME	30	3	66.7
WORK-OF-ART	20	8	0.0
LAW	6	0	NA
LANGUAGE	1	0	NA
PRODUCT	1	1	0.0
Micro-avg	—	—	53.8

Table 5: NER per-type performance on the test set when pre-trained on English data and fine-tuned on semi-automatic projected Chinese data.



(a) Semi-automatic vs unsupervised spans



(b) Semi-automatic vs silver spans

Figure 5: Disagreements of semi-automatic spans with automatic spans in NER.